EXHIBIT 28

Overfitting Regression Models: Problems, Detection, and Avoidance

May 26, 2017 By Jim Frost — 8 Comments

Overfitting a model is a condition where a statistical model begins to describe the random error in the data rather than the relationships between variables. This proble occurs when the model is too complex. In regression analysis, overfitting can produc misleading R-squared values, regression coefficients, and p-values. In this post, I exp how overfitting models is a problem and how you can identify and avoid it.

Overfit regression models have too many terms for the number of observations. Whe this occurs, the regression coefficients represent the noise rather than the genuine relationships in the population.

That's problematic by itself. However, there is another problem. Each sample has its unique quirks. Consequently, a regression model that becomes tailor-made to fit the random quirks of one sample is unlikely to fit the random quirks of another sample. Thus, overfitting a regression model reduces its generalizability outside the original dataset.

Taking the above in combination, an overfit regression model describes the noise, and it's not applicable outside the sample. That's not very helpful, right? I'd really like

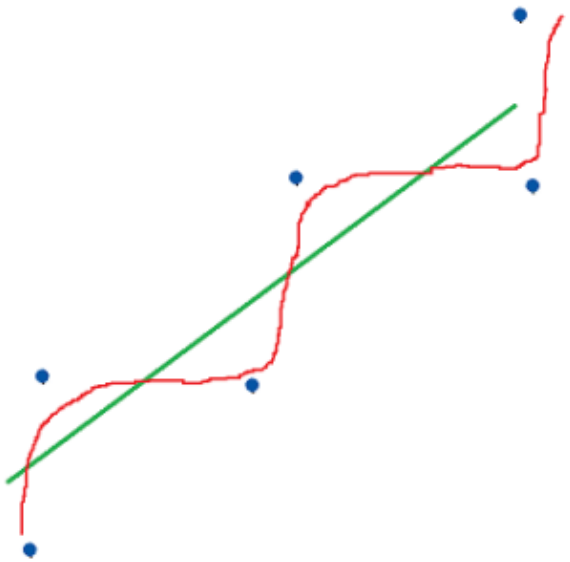
these problems to sink in because overfitting often occurs when analysts chase a high R-squared. In fact, inflated R-squared values are a *symptom* of overfit models! Despite the misleading results, it can be difficult for analysts to give up that nice high R-squared value.

When choosing a regression model, our goal is to approximate the true model for the whole population. If we accomplish this goal, our model should fit most random samples drawn from that population. In other words, our results are more generalizable—we can expect that the model will fit other samples.

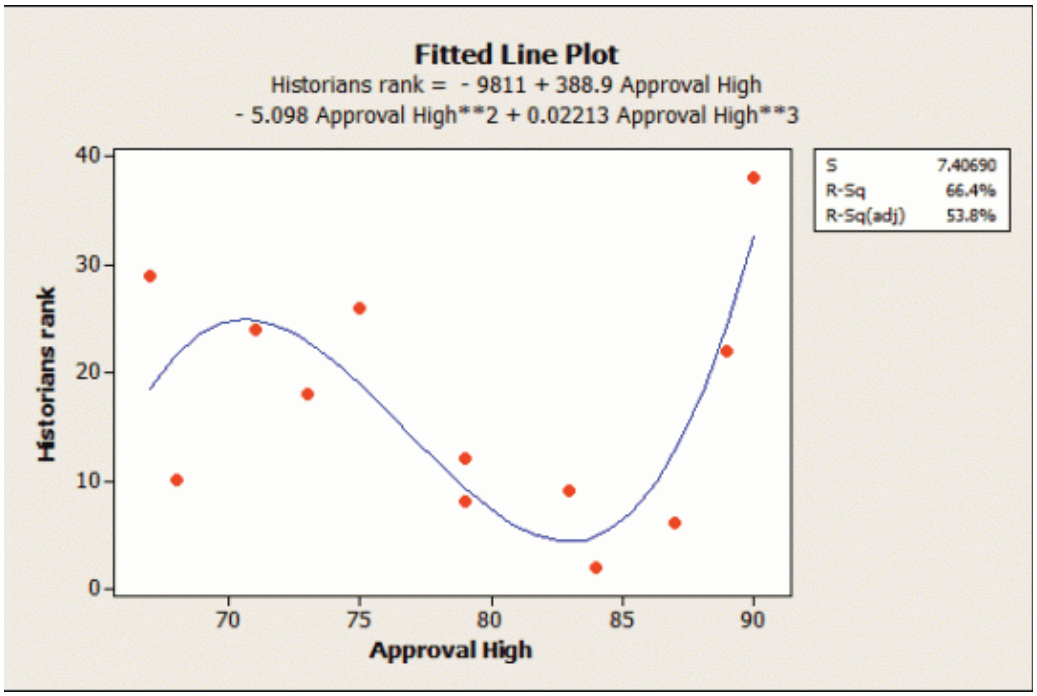
Related post: Model Specification: Choosing the Correct Regression Model

Graphical Illustration of Overfitting Regression Models

The image below illustrates an overfit model. The green line represents the true relationship between the variables. The random error inherent in the data causes the data points to fall randomly around the green fit line. The red line represents an overfit model. This model is too complex, and it attempts to explain the random error present in the data.



The example above is very clear. However, it's not always that obvious. Below, the fitt line plot shows an overfit model. In the graph, it appears that the model explains a g proportion of the dependent variable variance. Unfortunately, this is an overfit mode and I'll show you how to detect it shortly.



If you have more than two independent variables, it's not possible to graph them in t manner, which makes it harder to detect.

How Overfitting a Model Causes these Problems

Let's go back to the basics of inferential statistics to understand how overfitting mode causes problems. You use inferential statistics to draw conclusions about a populatio from a random sample. An important consideration is that the sample size limits the quantity and quality of the conclusions you can draw about a population. The more y need to learn, the larger the sample must be.

This concept is fairly intuitive. Suppose we have a total sample size of 20 and we nee estimate one population mean using a 1-sample t-test. We'll probably obtain a good estimate. However, if we want to use a 2-sample t-test to estimate the means of two populations, it's not as good because we have only ten observations to estimate each

mean. If we want to estimate three or more means using one-way ANOVA, it become pretty bad.

As the number of observations per estimate decreases (20, 10, 6.7, etc.), the estimate become more erratic. Furthermore, a new sample is unlikely to replicate the inconsis estimates produced by the smaller sample sizes.

In short, the quality of the estimates deteriorates as you draw more conclusions from sample. This idea is directly related to the degrees of freedom in the analysis. To lear more about this concept, read my post: Degrees of Freedom in Statistics.

Applying These Concepts to Overfitting Regression Models

Overfitting a regression model is similar to the example above. The problems occur when you try to estimate too many parameters from the sample. Each term in the m forces the regression analysis to estimate a parameter using a fixed sample size. Therefore, the size of your sample restricts the number of terms that you can safely a to the model before you obtain erratic estimates.

Similar to the example with the means, you need a sufficient number of observations each term in the regression model to help ensure trustworthy results. Statisticians ha conducted simulation studies* which indicate you should have at least 10-15 observations for each term in a linear model. The number of terms in a model is the of all the independent variables, their interactions, and polynomial terms to model curvature.

For instance, if the regression model has two independent variables and their interac term, you have three terms and need 30-45 observations. Although, if the model has multicollinearity or if the effect size is small, you might need more observations.

To obtain reliable results, you need a sample size that is large enough to handle the model complexity that your study requires. If your study calls for a complex model, y must collect a relatively large sample size. If the sample is too small, you can't

dependably fit a model that approaches the true model for your independent variabl
In that case, the results can be misleading.

How to Detect Overfit Models

As I discussed earlier, generalizability suffers in an overfit model. Consequently, you
detect overfitting by determining whether your model fits new data as well as it fits th
data used to estimate the model. In statistics, we call this cross-validation, and it ofte
involves partitioning your data.

However, for linear regression, there is an excellent accelerated cross-validation met
called predicted R-squared. This method doesn't require you to collect a separate sar
or partition your data, and you can obtain the cross-validated results as you fit the
model. Statistical software calculates predicted R-squared using the following automa
procedure:

- It removes a data point from the dataset.
- Calculates the regression equation.
- Evaluates how well the model predicts the missing observation.
- And, repeats this for all data points in the dataset.

Predicted R-squared has several cool features. First, you can just include it in the outp
as you fit the model without any extra steps on your part. Second, it's easy to interpre
You simply compare predicted R-squared to the regular R-squared and see if there is
big difference.

If there is a large discrepancy between the two values, your model doesn't predict ne
observations as well as it fits the original dataset. The results are not generalizable, a
there's a good chance you're overfitting the model.

For the fitted line plot above, the model produces a predicted R-squared (not shown)
0%, which reveals the overfitting. For more information, read my post about how to
interpret predicted R-squared, which also covers the model in the fitted line plot in m
detail.

How to Avoid Overfitting Models

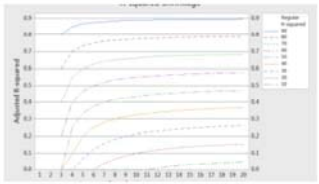
To avoid overfitting a regression model, you should draw a random sample that is lar enough to handle all of the terms that you expect to include in your model. This proc requires that you investigate similar studies before you collect data. The goal is to identify relevant variables and terms that you are likely to include in your own model After you get a sense of the typical complexity of models in your study area, you'll be able to estimate a good sample size.

For more information about successful regression modeling, read my post: Five Regression Analysis Tips to Avoid Common Mistakes.

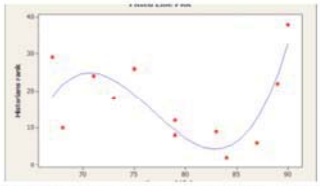
Reference

Babyak, MA., What You See May Not Be What You Get: A Brief, Nontechnical Introduc to Overfitting in Regression-Type Models, *Psychosomatic Medicine* 66:411-421 (2004)

Related Posts on Statistics by Jim



Five Reasons Why Your R-squared can be Too High
In "Regression"



How to Interpret Adjusted R-Squared and Predicted R-Squared in Regression Analysis
In "Regression"

Model Specification: Choosi the Correct Regression Moc
In "Regression"

Filed Under: Regression Tagged With: conce

Comments

Md Rabiul Kabir says
October 6, 2017 at 10:38 pm

Very helpful site

Reply

Jim Frost says
October 6, 2017 at 10:51 pm

Thank you! I'm glad you found it to be helpful!

Reply

Ramskrishna says
October 9, 2017 at 11:43 am

Wonderful job thank you

Reply

Jim Frost says
October 9, 2017 at 12:08 pm

Thanks so much for your kind comment. It made my day!

Reply

Ed says
January 25, 2018 at 9:57 am

I've been asked to right a proof for why the number of regressors K cannot exceed N. understand the intuition need some help proving it mathematically.

Reply

Jim Frost says
January 25, 2018 at 11:24 am

Hi Ed, here's a pointer in the right direction. When the number of parameters = N there are zero error degrees of freedom. Note that the parameters include the constant. So, if you have five observations, you can estimate the parameters for constant and four predictors.

Reply

reet khatri says
February 26, 2018 at 3:50 am

this is so easy to understand ,thank you

Reply

Jim Frost says
February 26, 2018 at 10:01 am

Hi Reet, you're very welcome! I'm happy to hear that you found it to be helpful!

Reply

Leave a Reply

Enter your comment here...

Meet Jim

I'll help you intuitively understand statistics by focusing on concepts and using plain English so you can concentrate on understanding your results.

Read More…

Search this website …

Subscribe via Email!

Enter your email address to receive notifications of new posts by email.

Email Address

Subscribe

Follow Me

Facebook

RSS Feed

Twitter

Popular

How To Interpret R-squared in Regression Analysis

How to Interpret P-values and Coefficients in Regression Analysis

How to Interpret the F-test of Overall Significance in Regression Analysis

Standard Error of the Regression vs. R-squared

Understanding Interaction Effects in Statistics

Latest